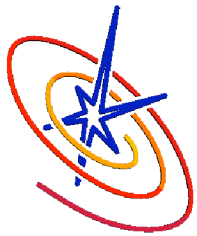


The Center for Astrophysical Thermonuclear Flashes

FLASH on BG/L

Rusty Lusk
Argonne National Laboratory

Andrew Siegel, Tomasz Plewa
University of Chicago

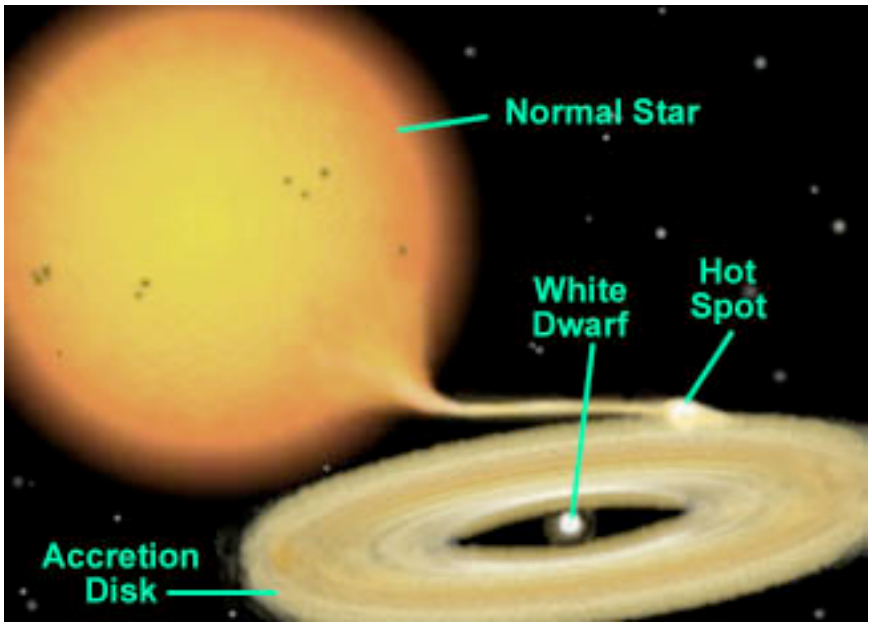
An Accelerated Strategic Computing Initiative (ASCI)
Academic Strategic Alliances Program (ASAP) Center
at The University of Chicago

Outline

- The Center for Astrophysical Thermonuclear Flashes at the University of Chicago
- The FLASH code
  - Functionality
  - Structure
- Science drivers for FLASH
- Scalability of FLASH (so far)
- Scalability analysis/tools
- Potential of FLASH on BG/L

FLASH Center Goals

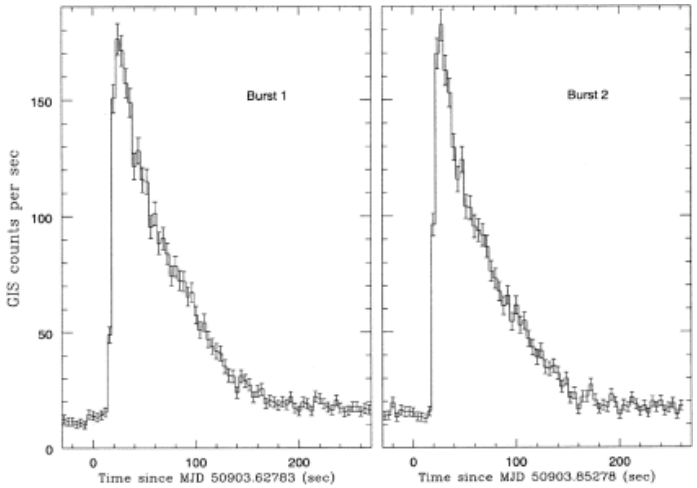
- To simulate matter accretion onto the surfaces of compact stars, nuclear ignition of the accumulated (and possibly stellar) material, and the subsequent evolution of the star's interior, surface, and exterior
  - Novae (on white dwarf surfaces)
  - Type Ia supernovae (in white dwarf interiors)
  - X-ray bursts (on neutron star surfaces)



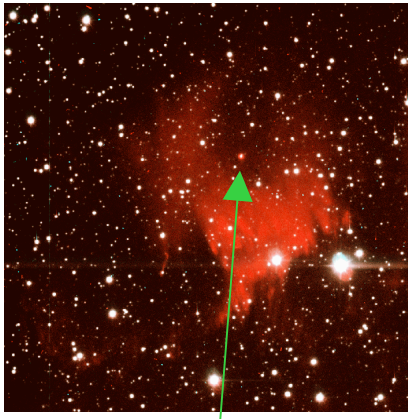
The FLASH Code

- Astrophysics code from the University of Chicago ASCI Alliance Center
  - Fluid dynamics, nuclear physics, MHD
  - Adaptive mesh refinement (Paramesh library does all MPI calls)
  - High quality
    - Thorough testing framework
    - Extensive documentation
    - Modular structure – easy to add more physics
- Principal communication patterns
  - ghost-cell exchanges (relatively local)
  - rebalancing grid after refinement
- Current scaling
  - studied extensively on multiple existing platforms
  - most physics scales to 1000's of procs
  - new algorithms need more analysis

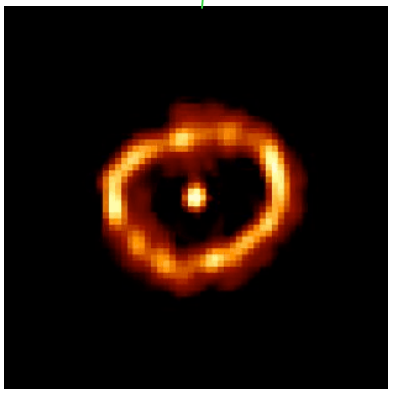
Observations of Astrophysical Flashes



X-ray burst: GS 1826-24



Nova: Nova Cygni 1992



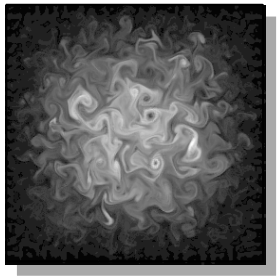
Credit: NASA/STScI

Type Ia SN: SN 1994D

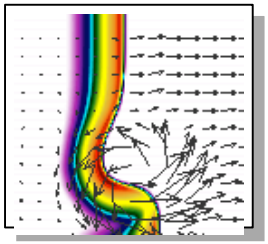
Credit: NASA/STScI

FLASH Scientific Results

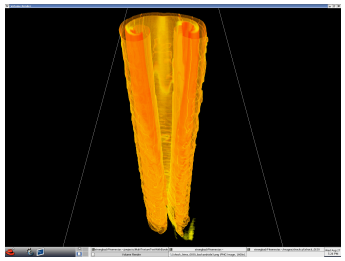
- Wide range of compressibility
- Wide range of length and time scales
- Many interacting physical processes
- Only indirect validation possible
- Rapidly evolving computing environment
- 2D → 3D



Compressible turbulence

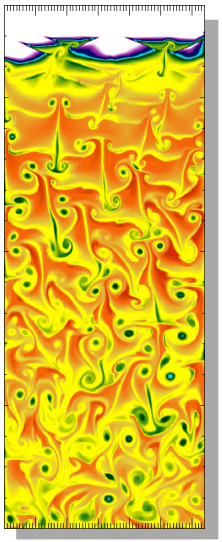


Flame-vortex interactions

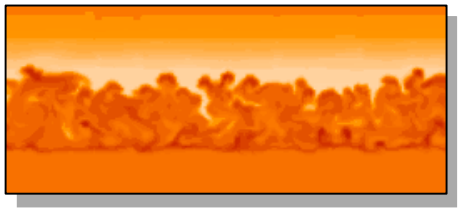


Shocked cylinder

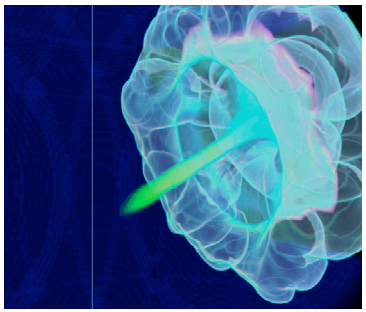
Richtmyer-Meshkov instability



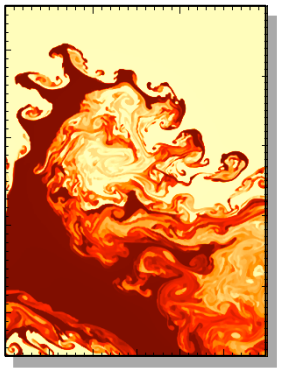
Cellular detonations



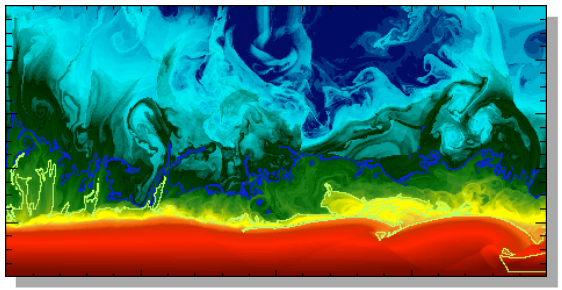
Nova outbursts on white dwarfs



White Dwarf deflagration



Rayleigh-Taylor instability



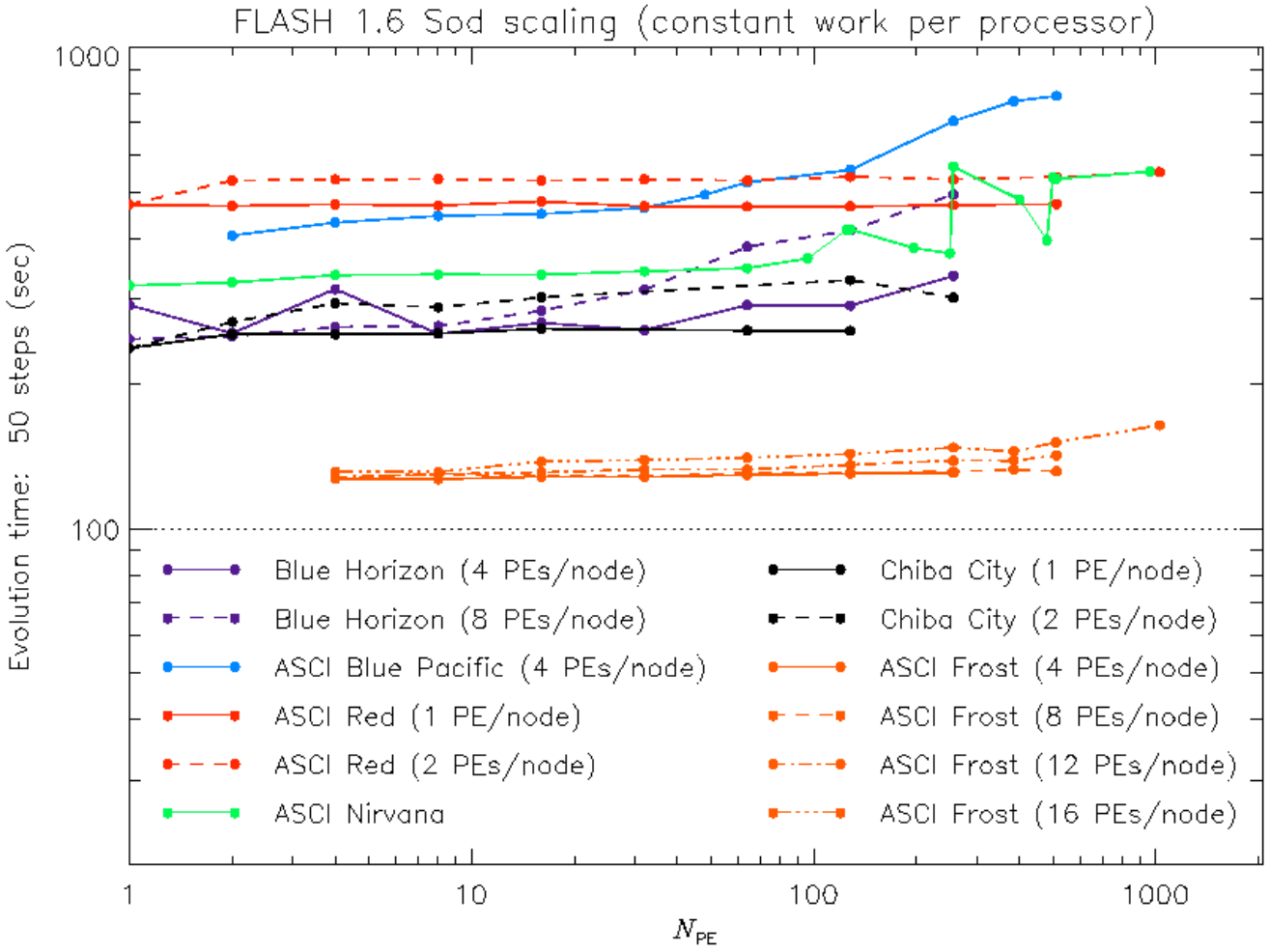
Helium burning on neutron stars

Scalability of Flash Algorithms

- Known scalable algorithms (all of these are AMR)
  - Explicit hydro
  - Equation of State
  - Nuclear physics
  - Parallel I/O
    - Using MPI-IO, HDF-5, Parallel NetCDF
- Scalability still being explored in some areas
  - Multipole gravity on AMR mesh
  - Multigrid solves on AMR mesh
    - Used for both self-gravity and implicit hydro
  - Some good preliminary results on each
  - Not all issues well understood yet

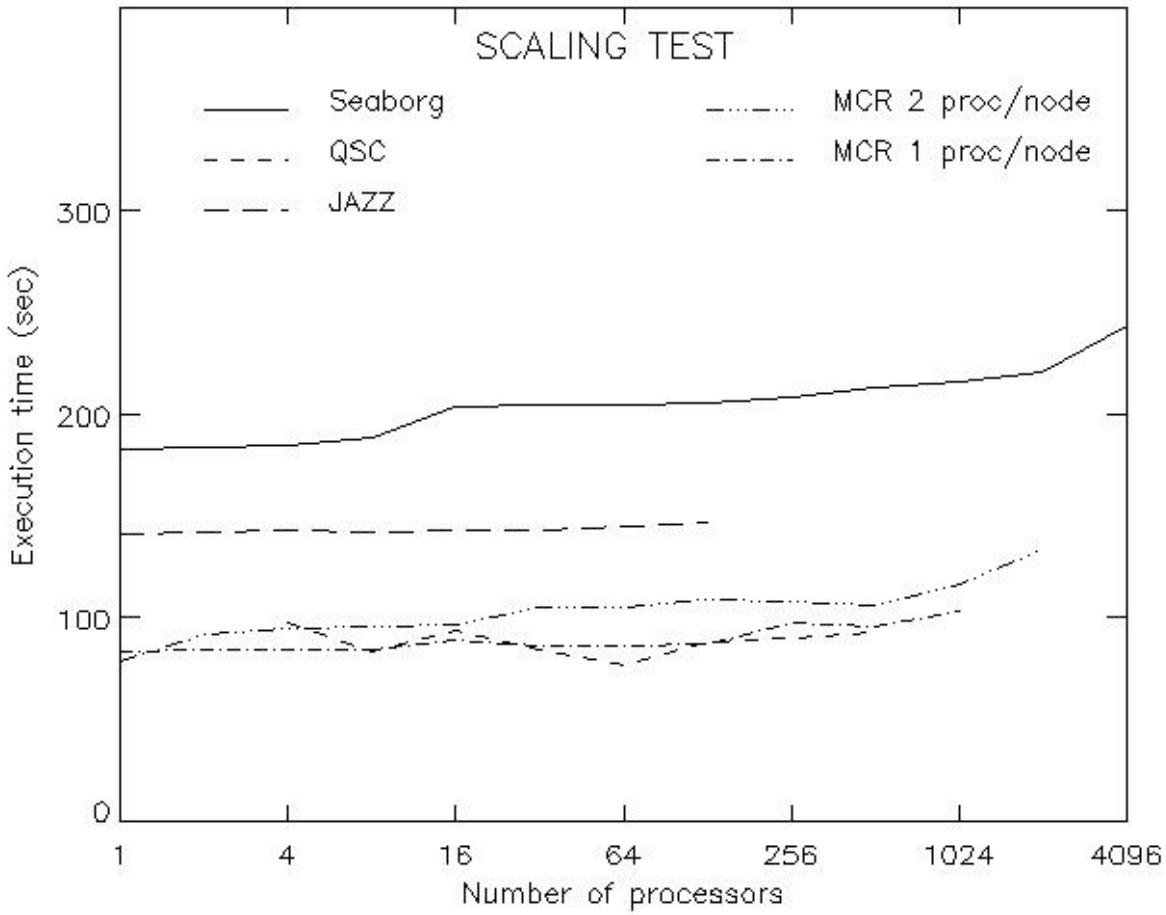
Scaling Results I

Scaling of Flash explicit physics on older ASCI machines



Scaling Results II

Scaling of explicit physics in FLASH2 on more recent machines



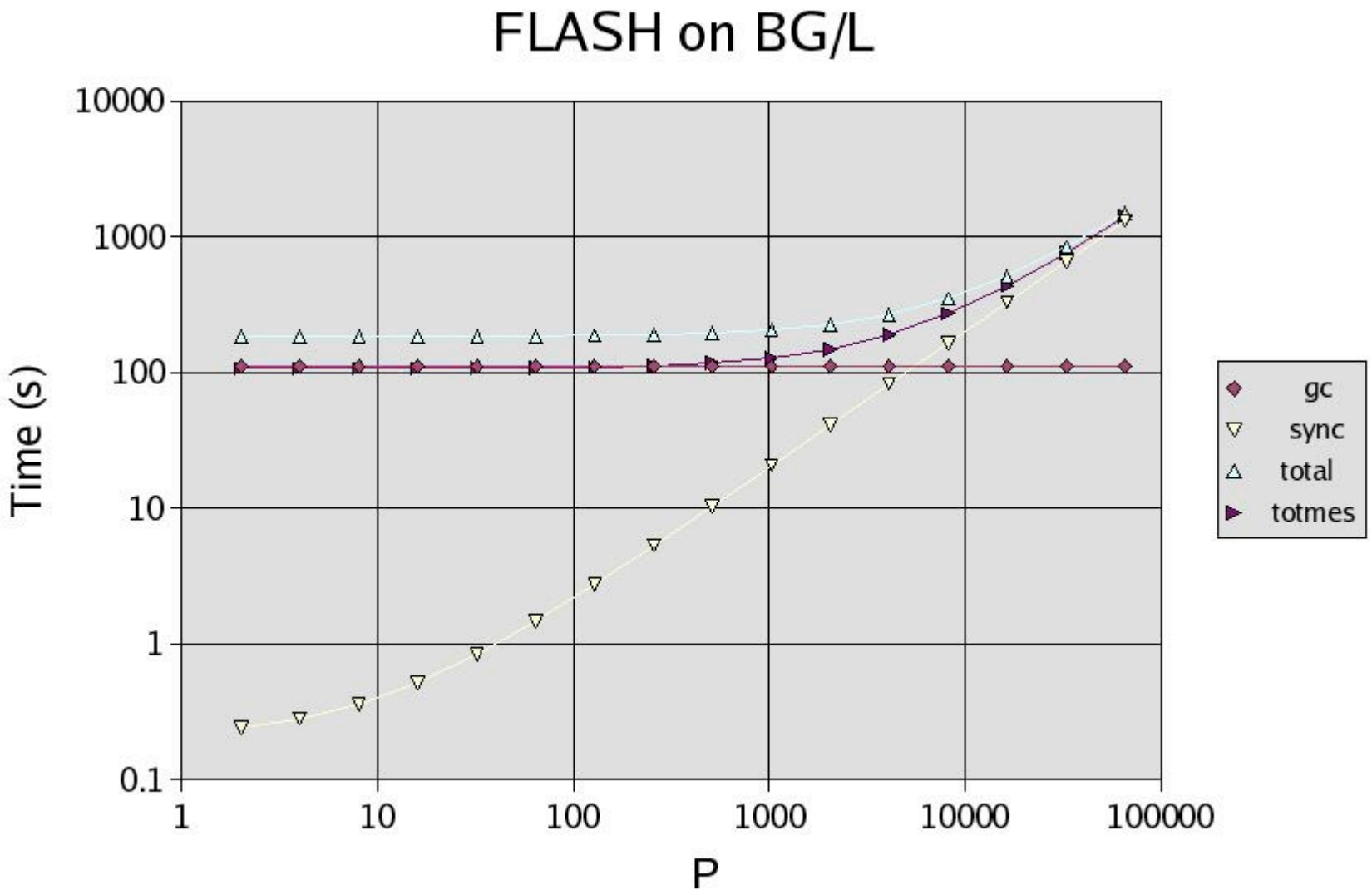
Anticipated Scalability Issues for FLASH on BG/L

- Block redistribution algorithm
  - Not a problem so far, but will be; solutions identified
- Enough memory/node to avoid starvation?
  - 256K/proc is enough, but more is better
  - Used FLASH performance model to demonstrate this
- Topology flat enough?
  - Research needed, but MPI point-to-point differences not *that* great, especially for long messages
  - We believe collective operations will be optimized by the MPI implementation
  - Plan to build into performance model
- How best to use the paired CPU's?
- Contention – load balancing issues at > 10K procs?
- So far MPI seems adequate, but perhaps alternate programming models will be more effective; e.g. multiple BG/L processes per MPI proc?

Tools to Help with Scalability Research

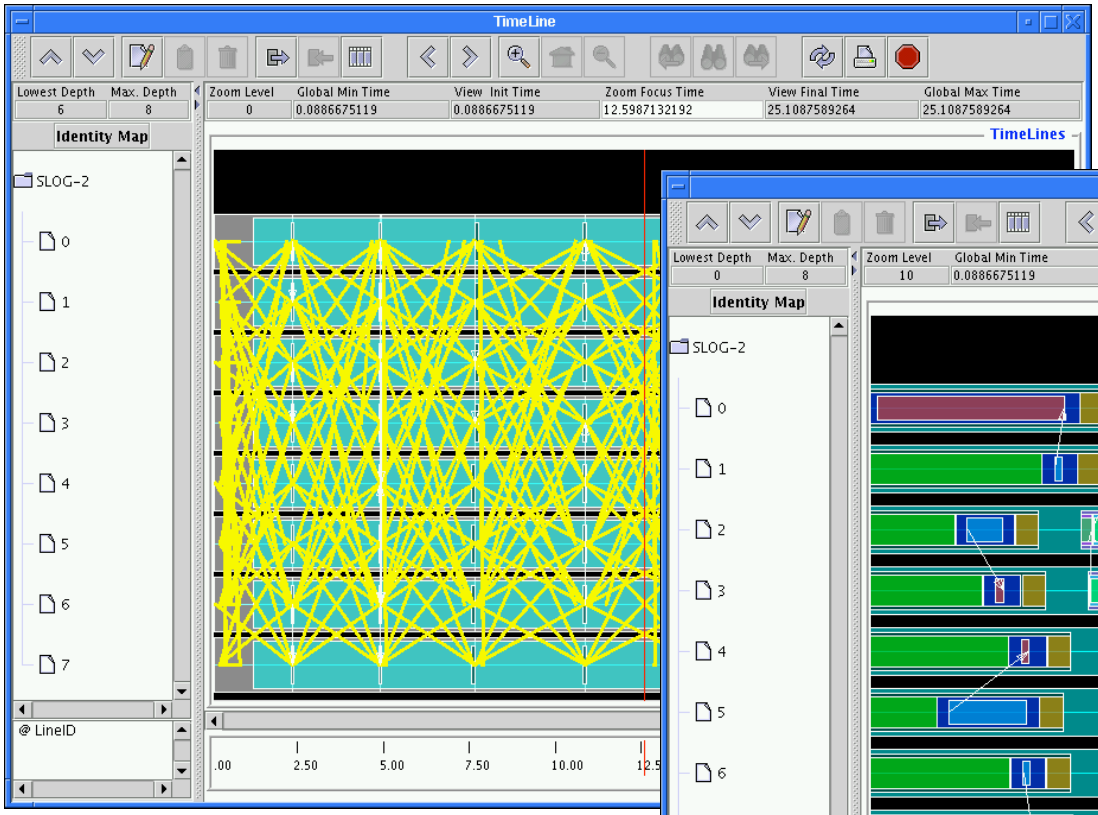
- FLASH Performance Modeler (Dursi & Riley)
  - Predicts parallel and single proc performance
    - Uses number of processors, message info, cache characteristics
  - Initially well verified on several platforms to 1000 proc
  - Current simplifications
    - 2-D
    - No multigrid/multipole
    - Typical model assumptions (cache behavior, etc.)
- FPMPI (Gropp)
  - Lightweight MPI profiling library
  - Can get average distance of mgs, msg size, histograms
  - Provides realistic input for FLASH Performance Modeler
- Jumpshot (Lusk & Chan)
  - Scalable logfiles; high-performance viewer
  - Collaboration with IBM (D. Wootten)
- Can explore MPI customizations with MPICH-2

Flash Performance Model Results



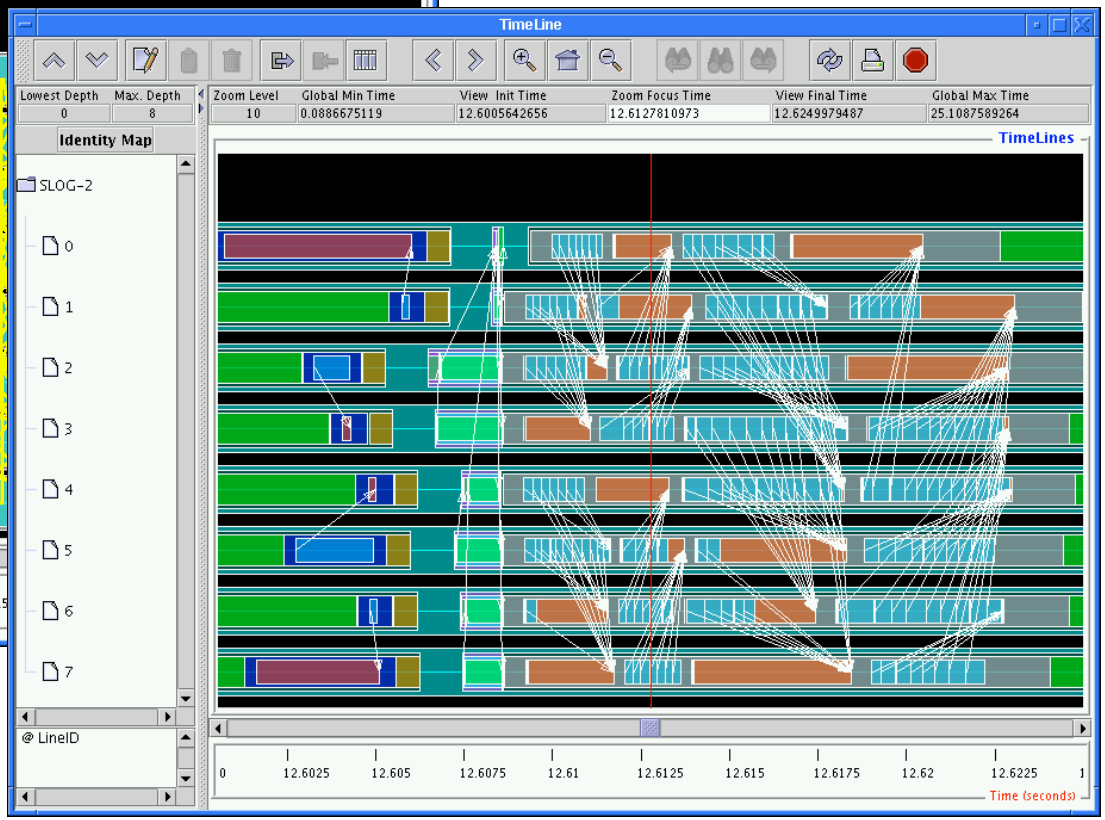
Initial verifications on Linux clusters

Using Jumpshot to look at FLASH at multiple time scales



Each line represents 1000_s of messages

1000 x



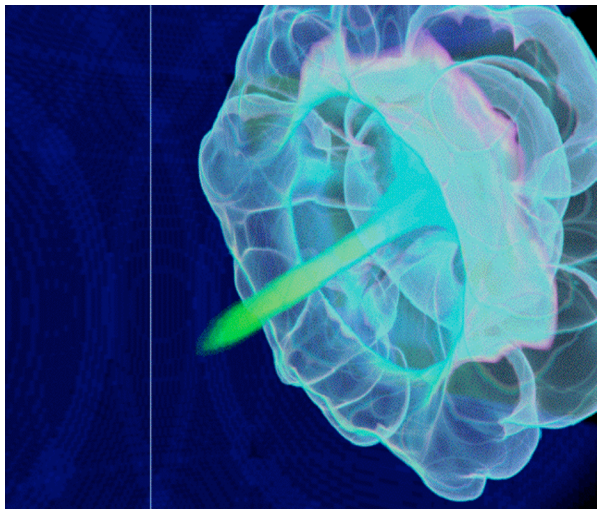
Detailed view shows opportunities for optimization

MPICH-2

- Portable, high-performance implementation of full MPI-2 standard
  - Open-source MPI implementation to promote MPI standard
  - Research vehicle for MPI implementation issues
  - Abstract Device Interface allows customization for specific networks
  - Process Management interface allows multiple process managers
- Starting point for multiple specialized MPI's
  - IBM's BG/L (collaboration with Jose Moreira and George Almasi)
  - Cray's Red Storm
  - Myricom's Myrinet Cluster
  - Ohio State U.'s Infiniband-based implementation (collab. with D.K. Panda)
- Recent Developments
  - All new collective operation implementations
    - Custom collective implementations can be used on a communicator basis
  - All new datatype handling now better than most "by hand" packing
- Status
  - Current release (0.95) lacks only passive target RMA and parts of dynamic at upper levels – full MPI-2 by SC03
  - Thread-safe at MPI_THREAD_FUNNELLED level
  - Shared-memory + TCP implementation of ADI
  - Remote-memory-access-based ADI implementation in progress

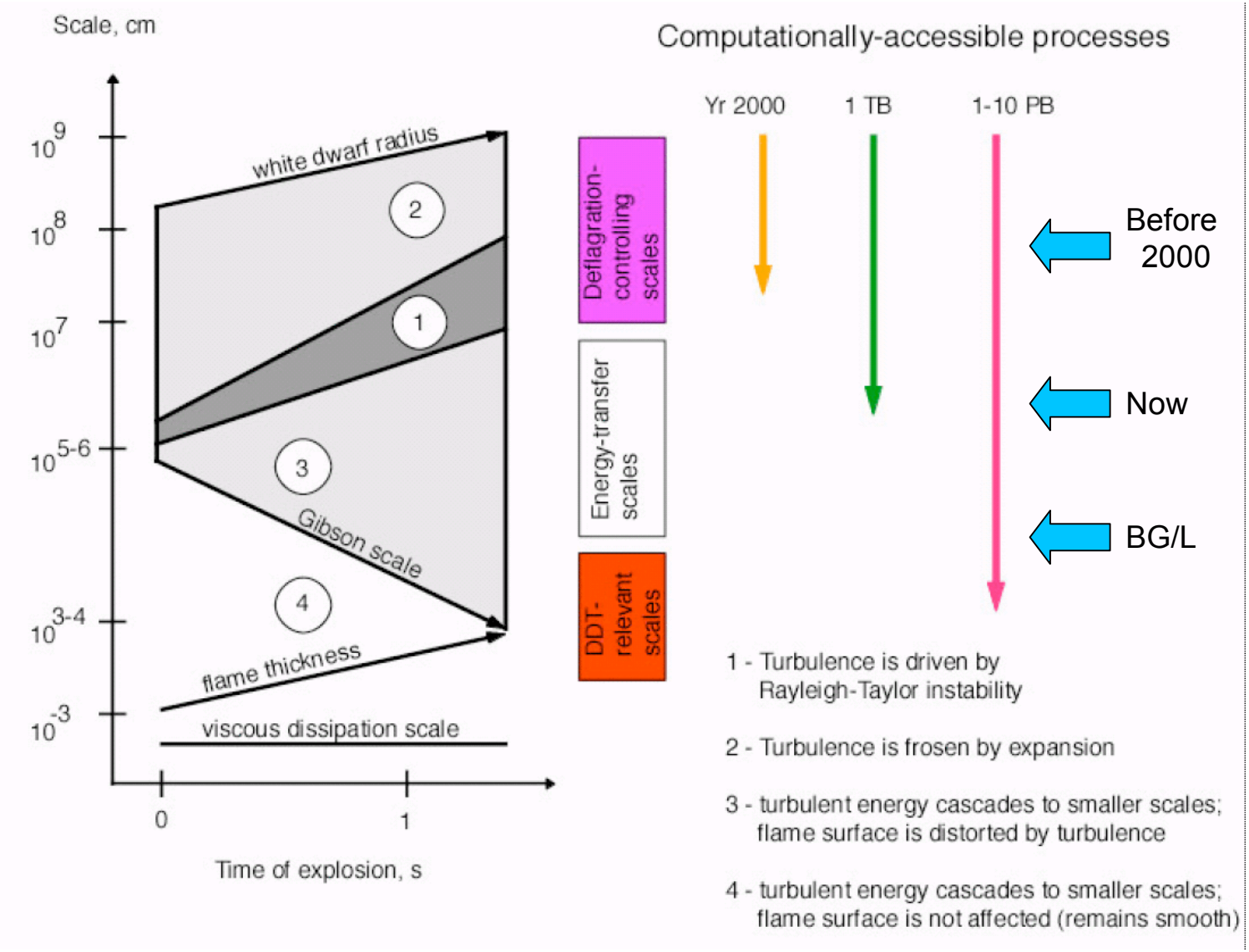
What would we do with 131,000 processors?

- Astrophysics is rich in large-scale scalable problems
  - Cosmology
    - Gravitational collapse
  - Star formation
    - Turbulence and gravitational collapse
  - Thermonuclear flames
    - Localized turbulence and nuclear reactions



- Best BG/L computational target:
  - Type Ia supernova explosion at 1km resolution
    - Major scientific achievement – unprecedented resolution
    - Most algorithms involved are likely to scale well in current form, or with only small modifications
    - Multipole solver scalability issues need to be researched further
    - 100K processes will also allow even more realistic physics

Length scales in Type Ia Supernova

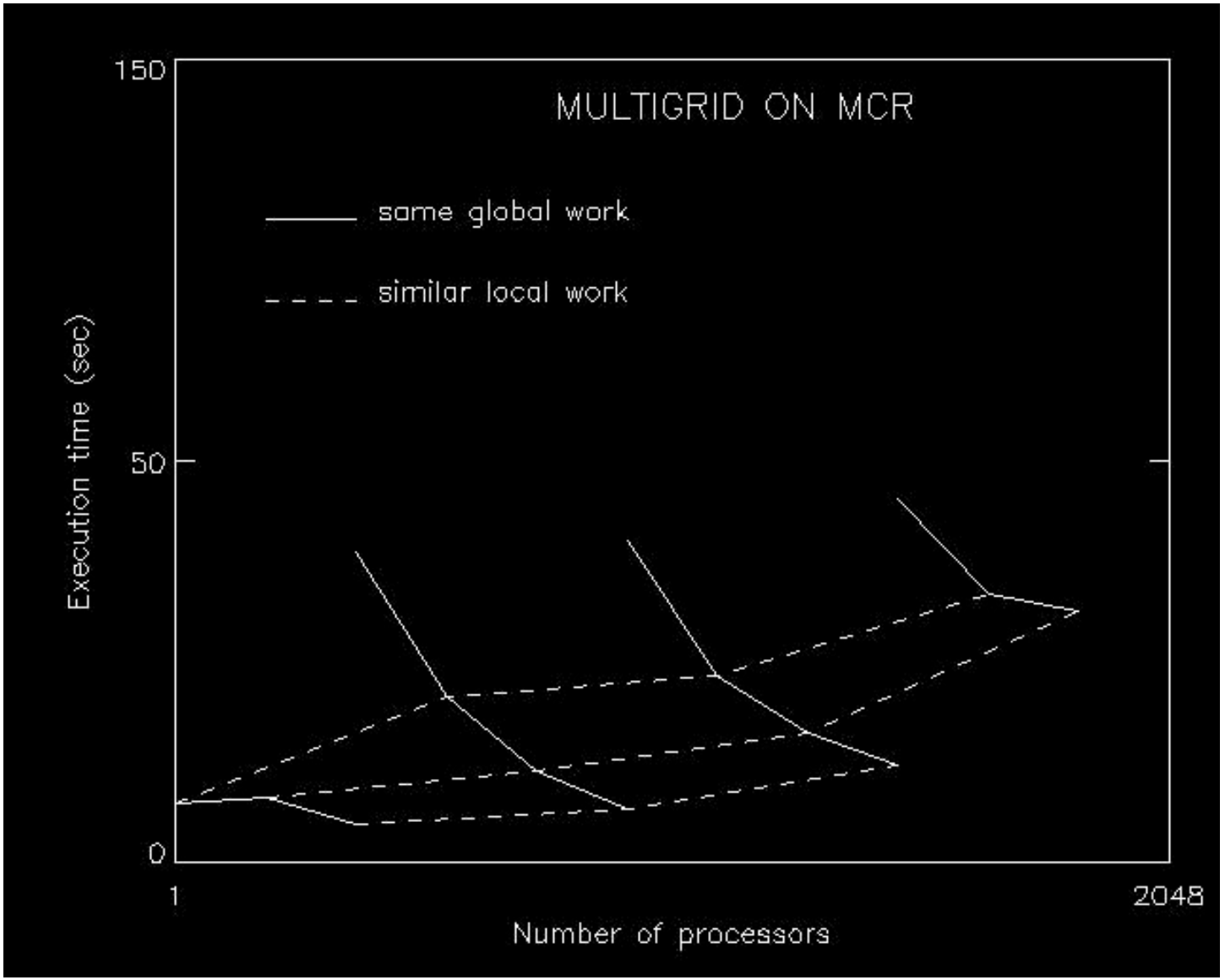


Summary

- The FLASH center is attacking problems of scientific importance that really require BG/L-type computing power.
- We are ready to do these problems now
- The question is: how will they perform on full BG/L machine?
- To address this question, we have done three things
  - carry out extensive scalability and FLOPS studies on a huge range of existing platforms
  - profile and reason about the scalability of our algorithms using FPMPI and Jumpshot
  - begin to model these algorithms for arbitrary machine parameters
- For most of our algorithms, we have identified several potential problem areas, but we think these are very surmountable
- The same is true for the implicit solves, but the work is more preliminary and more analysis is required.
- We are very excited about this opportunity!

The End

Scaling of multigrid solver -- MCR



Scaling of multigrid solver -- Seaborg